·论著·

心血管疾病中高风险人群颈动脉粥样硬化的识别：基于机器学习的预测模型及验证

刘忠典，许琪，陈伊静，覃玲巧，陈淑萍，唐薇婷，钟秋安*

530021 广西壮族自治区南宁市，广西医科大学公共卫生学院流行病学教研室

*通信作者：钟秋安，教授/博士生导师；E-mail：qazhong@gxmu.edu.cn

【摘要】 **背景** 颈动脉粥样硬化（CAS）常被视为心血管疾病（CVD）的预警信号，其诊断技术颈动脉多普勒超声检查没有被纳入公共卫生服务项目，同时弗雷明汉风险评分（FRS）存在着评估CAS风险准确性不足的情况，不利于基层医疗人员识别CAS。目前，关于机器学习方法识别FRS中高风险人群CAS的研究依然缺乏。**目的** 运用机器学习方法构建FRS中高风险人群CAS预测模型，比较其判别效能，筛选出性能最优的模型，以期辅助基层医疗人员更简便更准确地识别CAS。**方法** 选取2019—2021年和2023年在广西壮族自治区柳州市两乡镇符合纳排标准的674例当地居民作为研究对象。收集相关信息，并采集空腹血样、尿样检测生化指标。采用FRS评估CVD发生风险；运用颈动脉超声诊断CAS。将2019—2021年517例研究对象按照8∶2随机分为训练集和验证集，训练集用于构建Logistic回归、随机森林（RF）、支持向量机（SVM）、极端梯度增强（XGBoost）模型和梯度增强决策树（GBDT）模型，验证集用于内部验证；2023年157例研究对象作为测试集，用于外部验证。通过Lasso回归分析筛选特征变量，运用灵敏度、特异度、准确度、F1值和曲线下面积（AUC）值评价判别效能，外部验证采用AUC值评价最优模型泛化能力，并通过Shapley Additive exPlanation（SHAP）方法探讨影响最优模型识别CAS的重要变量。**结果** 通过Lasso回归，筛选出15个非零特征变量：年龄、BMI、收缩压（SBP）、吸烟、饮酒、高血压、总胆固醇、高密度脂蛋白胆固醇、C-反应蛋白（CRP）、空腹血糖、载脂蛋白B（ApoB）、脂蛋白a（LPA）、天冬氨酸氨基转移酶（AST）、AST/丙氨酸氨基转移酶、尿微量白蛋白肌酐比值。构建的Logistic回归、RF、SVM、XGBoost模型和GBDT模型的AUC值均较高，其中GBDT模型的判别性能最优，其灵敏度、特异度、准确度、F1值和AUC值分别是0.755 1、0.836 4、0.798 1、0.778 9、0.834 9，外部验证AUC值为0.794 0。SHAP方法发现年龄、SBP、CRP、LPA、ApoB是影响GBDT模型识别CAS排名前5的因素。**结论** 基于机器学习识别CAS的Logistic回归、RF、SVM、XGBoost模型和GBDT模型均显示出较高的判别性能，其中GBDT模型综合判别效能最佳，同时具有较强的泛化能力。

【关键词】 心血管疾病；颈动脉粥样硬化；机器学习；弗雷明汉风险评分；识别；预测

【中图分类号】 R 54　【文献标识码】 A　DOI：10.12114/j.issn.1007-9572.2024.0019

Identification of Carotid Atherosclerosis in Medium-high Risk Population of Cardiovascular Disease: Prediction Model and Validation Based on Machine Learning

*LIU Zhongdian, XU Qi, CHEN Yijing, QIN Lingqiao, CHEN Shuping, TANG Weiting, ZHONG Qiuan**

Department of Epidemiology, School of Public Health, Guangxi Medical University, Nanning 530021, China

**Corresponding author: ZHONG Qiuan, Professor/Doctoral supervisor; E-mail: qazhong@gxmu.edu.cn*

【Abstract】 **Background** Carotid atherosclerosis (CAS) is often considered an early warning signal for cardiovascular diseases (CVD). The diagnostic technique of carotid artery Doppler ultrasonography has not been included in public health service programs, and the Framingham Risk Score (FRS) lacks accuracy in assessing CAS risk, hindering the identification of CAS by primary healthcare personnel. Currently, there is a lack of research on machine learning methods to identify CAS in the medium-high risk population assessed by FRS. **Objective** To construct a CAS risk prediction model for the medium-high risk

基金项目：国家自然科学基金资助项目（82060088）

引用本文：刘忠典，许琪，陈伊静，等.心血管疾病中高风险人群颈动脉粥样硬化的识别：基于机器学习的预测模型及验证［J］.中国全科医学，2024. DOI：10.12114/j.issn.1007-9572.2024.0019.［Epub ahead of print］.［www.chinagp.net］

LIU Z D, XU Q, CHEN Y J, et al. Identification of carotid atherosclerosis in medium-high risk population of cardiovascular disease: prediction model and validation based on machine learning［J］. Chinese General Practice, 2024.［Epub ahead of print］.

population assessed by FRS using machine learning methods, compare its discriminative efficacy, select the optimal model, and assist primary healthcare personnel in identifying CAS more conveniently and accurately. **Methods** A total of 674 local residents from two townships in Liuzhou City, Guangxi Zhuang Autonomous Region, who met the inclusion criteria from 2019 to 2021 and 2023, were selected as the study subjects. Relevant information was collected, and biochemical indicators were measured in fasting blood and urine samples. FRS was used to assess the risk of CVD occurrence, and carotid ultrasound was used to diagnose CAS. Among the 517 subjects from 2019 to 2021, a random 8 ∶ 2 split was used to create a training set and a validation set. The training set was used to build Logistic regression, Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), and Gradient Boosting Decision Tree (GBDT) models, while the validation set was used for internal validation. The 157 subjects from 2023 served as the test set for external validation. Feature variables were selected using Lasso regression analysis, and discriminative efficacy was evaluated using sensitivity, specificity, accuracy, F1 score, and area under curve (AUC) value. External validation assessed the generalization ability of the optimal model using AUC value, and the Shapley Additive exPlanation (SHAP) method explored the important variables influencing the optimal model′s identification of CAS. **Results** Lasso regression analysis identified 15 feature variables: age, BMI, systolic blood pressure (SBP), smoking, drinking, hypertension, total cholesterol, high density lipoprotein cholesterol, C-reactive protein (CRP), fasting plasma glucose, apolipoprotein B (ApoB), lipoprotein (a) (LPA), aspartate aminotransferase (AST), AST/ alanine aminotransferase, urinary microalbumin creatinine ratio. The constructed Logistic regression, RF, SVM, XGBoost, and GBDT models exhibited high AUC values, with the GBDT model showing the best discriminative performance. Its sensitivity, specificity, accuracy, F1 score, and AUC value were 0.755 1, 0.836 4, 0.798 1, 0.778 9, and 0.834 9, respectively, and the external validation AUC value was 0.794 0. The SHAP method revealed that age, SBP, CRP, LPA, and ApoB were the top five factors influencing the GBDT model′s identification of CAS. **Conclusion** Logistic regression, RF, SVM, XGBoost, and GBDT models for identifying CAS based on machine learning all demonstrated high discriminative performance, with the GBDT model exhibiting the best comprehensive discriminative efficacy and strong generalization ability.

【Key words】 Cardiovascular diseases; Carotid atherosclerosis; Machine learning; Framingham risk score; Identification; Forecasting

心血管疾病（cardiovascular disease，CVD）是城乡居民主要死亡原因之一，其发病率和死亡率仍在不断上升，是中国居民的首要健康危险因素[1]。动脉粥样硬化是CVD的主要病理基础，其中颈动脉往往是最早受累的部位，因此，颈动脉粥样硬化（carotid atherosclerosis，CAS）通常被认为是CVD的预警信号[2]。在诊断方面，多普勒超声检测颈动脉内-中膜厚度（carotid intima-media thickness，CIMT）是判断CAS病变程度的可靠技术[3]。2009年以来，基本公共卫生服务项目在不断“扩容”，到2019年增加到12类服务项目[4]，但颈动脉多普勒超声检查并没有被纳入其中，不能满足CVD早期防治的需求；弗雷明汉风险评分（Framingham Risk Score，FRS）是被广泛应用的一种心血管风险评估方法，但其存在着评估CAS风险准确性不足的情况[5-6]，可能会导致基层医疗人员不能准确识别CAS。因此，亟需探索更简便有效的方法以早期识别CAS。近年来越来越多学者采用机器学习通过容易获取的因素对疾病进行识别，在个体自测和临床应用上均取得良好的效果[7]。

目前，关于机器学习识别FRS中高风险群体CAS的研究报道相对较少，为加强这一方面的研究，本研究选用Logistic回归、随机森林（Random Forest，RF）、支持向量机（Support Vector Machine，SVM）、极端梯度增强（Extreme Gradient Boosting，XGBoost）和梯度增强决策树（Gradient Boosting Decision Tree，GBDT）构建FRS中高风险群体（FRS>6%）CAS预测模型，并筛选出最优模型，以期辅助基层医疗人员更简便、更准确、更早地识别CAS，为临床防治工作提供科学依据。

1 对象与方法

1.1 研究对象

采用方便抽样法，于2019—2021年和2023年在广西壮族自治区柳州市两个乡镇选取当地居民1 169例作为研究对象，其中2019—2021年852例居民用于模型构建及内部验证，2023年317例居民用于外部验证。纳入标准：（1）30~74岁；（2）FRS>6%；（3）接受颈动脉多普勒超声检查。排除标准：（1）患有重大疾病的个体，如恶性肿瘤、严重感染性疾病、精神疾病等；（2）已被确诊为冠心病、脑卒中或外周动脉疾病；（3）协变量存在缺失。基于纳排标准，最终纳入674例（2019—2021年：517例；2023年：157例）符合条件的研究对象。本研究经广西医科大学伦理委员会批准（2019-SB-094），研究对象均已签署知情同意书。

1.2 研究方法

1.2.1 一般资料：通过课题组自行设计的问卷收集研究对象的性别、年龄、民族、受教育程度、体力活动、吸烟史、饮酒史、疾病史和药物使用情况等。体格检查主要包括BMI、腰围、心率、收缩压（SBP）及舒张压（DBP）。实验室检查指标包括总胆固醇（TC）、三酰甘油（TG）、低密度脂蛋白胆固醇（LDL-C）、高密度脂蛋白胆固醇（HDL-C）、空腹血糖（FPG）、尿微量白蛋白（ALB）、C-反应蛋白（CRP）、尿肌酐（UCR）、脂蛋白a（LPA）、载脂蛋白A（ApoA）、载脂蛋白B（ApoB）、丙氨酸氨基转移酶（ALT）及天冬氨酸氨基转移酶（AST），并计算尿微量白蛋白肌酐比值（ACR）=ALB/UCR。体力活动按国际体力活动问卷（短卷）[8]计算体力活动当量，以代谢当量（MET-min/w）表示。

1.2.2 FRS标准：本研究使用FRS评估人群CVD风险，将FRS>6%定义为CVD中高风险[9]。

1.2.3 CAS诊断：CAS定义为CIMT增加≥1 mm或斑块形成[10]。CIMT的定义及详细测量方法详见先前的研究[11]。斑块定义为侵犯动脉管腔至少0.5 mm或周围CIMT值的50%的局灶性结构，或CIMT>1.5 mm[12]。由专业的超声医师负责颈动脉多普勒超声检查，经专业化培训的调查人员负责相应数据的记录。根据CAS诊断结果将517例居民分为两组：正常组（272例）和CAS组（245例）。

1.2.4 相关定义：（1）吸烟，从未吸烟为总吸烟量<100支；曾经吸烟为>100支但调查前30 d未吸烟；当前吸烟为>100支且调查前30 d吸烟[13]。（2）饮酒，从未饮酒为饮酒<12个标准饮酒单位；曾经饮酒为既往饮酒≥12个标准饮酒单位但最近1年饮酒<1个标准饮酒单位；当前饮酒为既往饮酒≥12个标准饮酒单位且最近1年饮酒≥1个标准饮酒单位及以上[14]。（3）肾功能按慢性肾脏病流行病学协作公式计算估算肾小球滤过率（estimated glomerular filtration rate，eGFR），eGFR≥90 $mL\cdot min^{-1}\cdot(1.73\ m^2)^{-1}$定义为肾功能正常；eGFR<90 $mL\cdot min^{-1}\cdot(1.73\ m^2)^{-1}$定义为肾功能下降[15]。（4）高血压：参照《中国高血压防治指南（2018年修订版）》，SBP≥140 mmHg（1 mmHg=0.133 kPa）和/或DBP≥90 mmHg、既往诊断为高血压或正在服用降压药物者[16]。（5）糖尿病定义为本次调查FPG≥7.0 mmol/L，或自述有正在服用降糖药或患有糖尿病[17]。（6）脂代谢异常，TC≥200 mg/dL、TG≥150 mg/dL、LDL-C≥130 mg/dL、HDL-C<40 mg/dL、正在使用降脂药物，满足任意1项[18-19]。（7）代谢综合征依据国际糖尿病联盟对代谢综合征的定义[20]。（8）疾病一级亲属家族史，一级亲属（父亲、母亲、兄弟姐妹、儿子、女儿）中至少有1人患该疾病[11]。

1.3 模型构建

运用Python 3.7.4的scikit-learn 2.2.2库构建模型。将Lasso回归筛选出来的特征变量（连续变量进行归一化处理）作为输入变量，以CAS作为结局变量，使用scikit-learn 2.2.2中train_test_split模块将全部样本按照8∶2随机分为训练集和验证集，并保持划分后的数据集中阳性和阴性病例之间比例与全部数据集中的一致，在训练集中分别使用Logistic Regression、Random Forest Classifier、SVC、XGBClassifier、Gradient Boosting Classifier模块构建Logistic回归、RF、SVM、XGBoost模型和GBDT模型；采用GridSearchCV模块（网格搜索算法）对每个模型进行参数调优，将曲线下面积（area under curve，AUC）值作为评价指标。在验证集中采用灵敏度、特异性、准确度、F1值、AUC值评估5种模型的判别性能，筛选最优模型。在测试集中对最优模型进行外部验证，采用AUC值评估模型的泛化能力。

使用Shapley Additive exPlanation（SHAP）方法探讨每个特征变量对最优预测模型的具体影响。

1.4 统计学方法

采用R（4.1.3）进行统计分析。计量资料符合正态分布的以（$\bar{x}\pm s$）表示，两组间比较采用独立样本t检验；计量资料不符合正态分布的以M（P_{25}，P_{75}）表示，两组间比较采用Mann-Whitney U检验；计数资料以相对数表示，两组间比较采用χ^2检验。以CAS为因变量，使用Lasso回归分析筛选特征变量。绘制各模型在验证集中识别CAS的受试者工作特征（ROC）曲线，计算各模型的AUC值并比较。以P<0.05为差异有统计学意义。

2 结果

2.1 一般资料

517例居民中男210例（40.6%）、女307例（59.4%），平均年龄（60.2±7.9），CAS诊断245例（47.4%）、正常诊断272例（52.6%）。两组间性别、民族、受教育程度、腰围、心率、DBP、吸烟史、饮酒史、糖尿病、体力活动、FPG、TC、LDL-C、脂代谢异常、代谢综合征、ApoA、ApoB、ApoA/ApoB、ALT、AST、UCR、ALB、ACR比较，差异均无统计学意义（P>0.05）；两组间年龄、BMI、SBP、高血压、HDL-C、TG、肾功能、CRP、LPA、AST/ALT比较，差异均有统计学意义（P<0.05），见表1。

2.2 Lasso回归筛选特征变量

以居民是否诊断为CAS为因变量，以36个可能的影响因素为自变量进行Lasso回归模型筛选变量。其中分类变量赋值表见表2，年龄、心率、腰围、BMI、体力活动、SBP、DBP、FPG、TC、HDL-C、LDL-C、TG、CRP、LPA、ApoA、ApoB、ApoA/ApoB、UCR、



表 1 两组研究对象一般资料比较
Table 1 Comparison of general information between two groups

组别	例数	性别[例(%)] 男	性别[例(%)] 女	年龄($\bar{x}\pm s$,岁)	民族[例(%)] 汉族	民族[例(%)] 其他民族	受教育程度[例(%)] 初中及以下	受教育程度[例(%)] 高中及以上	BMI($\bar{x}\pm s$, kg/m^2)	腰围($\bar{x}\pm s$, cm)	心率[M(P_{25}, P_{75}), bpm]
正常组	272	108(39.7)	164(60.3)	56.9±7.2	182(66.9)	90(33.1)	247(90.8)	25(9.2)	24.5±3.2	84.8±8.7	79.0(71.0, 89.0)
CAS 组	245	102(41.6)	143(58.4)	63.9±7.1	158(64.5)	87(35.5)	225(91.8)	20(8.2)	23.5±3.3	83.4±9.3	80.0(72.0, 88.0)
检验统计量值		0.198[a]		−11.145[b]	0.237[a]		0.171[a]		3.544[b]	1.755[b]	−0.780
P 值		0.656		<0.001	0.627		0.679		<0.001	0.080	0.435

组别	SBP[M(P_{25}, P_{75}), mmHg]	DBP($\bar{x}\pm s$, mmHg)	吸烟史[例(%)] 从不吸烟	吸烟史[例(%)] 曾经吸烟	吸烟史[例(%)] 当前吸烟	饮酒史[例(%)] 从不饮酒	饮酒史[例(%)] 曾经饮酒	饮酒史[例(%)] 当前饮酒	高血压[例(%)] 否	高血压[例(%)] 是	糖尿病[例(%)] 否	糖尿病[例(%)] 是
正常组	140(130, 153)	83±12	196(72.1)	32(11.8)	44(16.1)	135(49.6)	44(16.2)	93(34.2)	112(41.2)	160(58.8)	230(84.6)	42(15.4)
CAS 组	147(134, 162)	82±13	161(65.7)	37(15.1)	47(19.2)	125(51.0)	35(14.3)	85(34.7)	76(31.0)	169(69.0)	211(86.1)	34(13.9)
检验统计量值	−3.644	0.998[b]	2.489[a]			0.360[a]			5.745[a]		0.251[a]	
P 值	<0.001	0.319	0.288			0.835			0.017		0.616	

组别	体力活动[M(P_{25}, P_{75}), MET-min/w]	FPG[M(P_{25}, P_{75}), mmol]	TC($\bar{x}\pm s$, mg/dL)	HDL-C[M(P_{25}, P_{75}), mg/dL]	LDL-C[M(P_{25}, P_{75}), mg/dL]	TG[M(P_{25}, P_{75}), mg/dL]	脂代谢异常[例(%)] 否	脂代谢异常[例(%)] 是
正常组	5 643.0(3 707.3, 8 325.8)	5.58(5.2, 6.1)	213.1±41.0	54.5(47.1, 63.8)	118.9(96.1, 139.3)	161.7(114.9, 222.8)	55(20.2)	217(79.8)
CAS 组	5 418.0(3 360.0, 9 039.0)	5.59(5.3, 6.3)	215.2±38.2	56.1(48.7, 69.6)	119.9(97.1, 141.9)	145.3(98.3, 194.9)	54(22.0)	191(78.0)
检验统计量值	−0.080	−1.097	−0.622[b]	−2.243	−0.409	−2.528	0.257[a]	
P 值	0.937	0.273	0.534	0.025	0.683	0.011	0.612	

组别	代谢综合征[例(%)] 否	代谢综合征[例(%)] 是	肾功能[例(%)] 正常	肾功能[例(%)] 下降	CRP[M(P_{25}, P_{75}), mg/L]	ApoA[M(P_{25}, P_{75}), g/L]	ApoB($\bar{x}\pm s$, g/L)	ApoA/ApoB[M(P_{25}, P_{75})]
正常组	146(53.7)	126(46.3)	132(48.5)	140(51.5)	1.0(0.5, 2.1)	1.6(1.5, 1.8)	1.1±0.3	1.4(1.1, 1.8)
CAS 组	143(58.4)	102(41.6)	80(32.7)	165(67.4)	1.3(0.7, 2.6)	1.7(1.5, 1.9)	1.1±0.3	1.5(1.2, 1.9)
检验统计量值	1.151[a]		13.430[a]		−2.878	−1.234	−0.148[b]	−0.621
P 值	0.283		<0.001		0.004	0.217	0.883	0.535

组别	LPA[M(P_{25}, P_{75}), g/L]	ALT[M(P_{25}, P_{75}), U/L]	AST[M(P_{25}, P_{75}), U/L]	AST/ALT[M(P_{25}, P_{75})]	UCR[M(P_{25}, P_{75}), g/L]	ALB[M(M(P_{25}, P_{75}), mg/L]	ACR[M(P_{25}, P_{75}), mg/g]
正常组	126.7(63.9, 244.3)	13.0(9.4, 19.0)	24.6(21.4, 30.0)	1.9(1.4, 2.5)	1.3(0.9, 1.7)	11.7(5.2, 22.7)	8.9(4.4, 17.6)
CAS 组	166.2(74.0, 322.0)	12.3(9.0, 17.6)	25.5(22.0, 31.9)	2.0(1.6, 2.6)	1.2(0.9, 1.6)	12.2(6.1, 27.9)	9.5(5.1, 22.8)
检验统计量值	−2.584	−1.151	−1.612	−2.561	−0.653	−1.179	−1.281
P 值	0.010	0.250	0.107	0.010	0.514	0.239	0.200

注：CAS= 颈动脉粥样硬化，SBP= 收缩压，DBP= 舒张压，FPG= 空腹血糖，TC= 总胆固醇，HDL-C= 高密度脂蛋白胆固醇，LDL-C= 低密度脂蛋白胆固醇，TG= 三酰甘油，CRP=C- 反应蛋白，ApoA= 载脂蛋白 A，ApoB= 载脂蛋白 B，ApoA/ApoB= 载脂蛋白 A/ 载蛋白 B，LPA= 脂蛋白 a，ALT= 丙氨酸氨基转移酶，AST= 天冬氨酸氨基转移酶，AST/ALT= 天冬氨酸氨基转移酶 / 丙氨酸氨基转移酶，UCR= 尿肌酐，ALB= 尿微量白蛋白，ACR= 尿微量白蛋白肌酐比值；[a] 为 χ^2 检验，[b] 为 t 值，余检验统计量值为 Z 值；1 mmHg=0.133 kPa。

ALB、ALT、AST、AST/ALT、ACR 均为实测值。最终筛选出 15 个非零系数变量：年龄、BMI、SBP、吸烟、饮酒、高血压、TC、HDL-C、CRP、FPG、ApoB、LPA、AST、AST/ALT、ACR（图 1 和表 3）。

2.3 构建机器学习模型

将 Lasso 回归筛选出的变量纳入 Logistic 回归、RF、SVM、XGBoost 模型及 GBDT 模型，通过网格搜索方法，以 AUC 值作为评价指标，在训练集中确定每个模型的最优参数分别为 Logistic 回归：solver=“liblinear”，max_iter=500，penalty=“l2”；RF：n_estimators=500，criterion=“gini”，bootstrap=True，max_depth=20，max_features=“auto”，min_samples_leaf=2，min_samples_split=2；SVM：kernel=“rbf”，C=1，gamma=0.01；XGBoost 模型：learning_rate=0.007，n_estimators=500，max_depth=2，min_child_weight=8，gamma=0.8，subsample=0.8，colsample_bytree=0.8，objective=“binary:logistic”，nthread=4；GBDT 模型：n_estimators=500，learning_rate=0.008，max_depth=2，subsample=0.8，max_features=“sqrt”，min_samples_split=5，min_samples_leaf=2，random_state=1117。

2.4 各模型对 CAS 的判别性能比较

将构建的 Logistic 回归、RF、SVM、XGBoost 模型和 GBDT 模型在验证集中进行内部验证，结果显示各模

型的 AUC 值均较高。其中 SVM 的 AUC 值最高，GBDT 模型灵敏度、特异度、准确度和 F1 值均最高。综合评价，GBDT 模型的判别性能最优，见图 2、表 4。

2.5 模型外部验证

对内部验证中性能最佳的 GBDT 模型进行外部验证，测试模型的泛化能力。结果显示，GBDT 模型外部验证集中的 AUC（0.794 0）较内部验证集（0.834 9）有所下降，但仍 >0.7，提示本研究构建的 GBDT 模型具有较强的外部泛化能力。

2.6 SHAP 方法探讨最优模型的解释性

在图 3A 中，按平均绝对 SHAP 值排序，展示了影响模型识别 CAS 的因素。这有助于直观地理解每个因素对模型识别的贡献程度。在图 3B 中，y 轴显示了每个变量的重要性，最重要的变量位于图表顶部，最不重要的变量位于底部。x 轴表示 SHAP 值，衡量了每个变量对模型识别的贡献大小。正值表示增加识别结果的可能性，负值表示减小识别结果的可能性。通过这个图表，可以清晰地了解每个变量对 CAS 识别的影响。点的颜色代表了变量的原始值，红色代表高值，蓝色代表低值。这就可以直观地观察变量的原始值与其对模型识别的影

表 2 Lasso 回归候选变量赋值表

Table 2 Lasso regression candidate variable assignment table

变量	赋值
CAS	否 =0，是 =1
性别	男性 =1，女性 =2
民族	汉族 =1，其他民族 =2
受教育程度	初中及以下 =1，高中及以上 =2
吸烟史	从不吸烟 =1，曾经吸烟 =2，当前吸烟 =3
饮酒史	从不饮酒 =1，曾经饮酒 =2，当前饮酒 =3
高血压	否 =0，是 =1
糖尿病	否 =0，是 =1
代谢综合征	否 =0，是 =1
脂代谢异常	否 =0，是 =1
肾功能	正常 =1，下降 =2
高血压一级亲属家族史	否 =0，是 =1
冠心病一级亲属家族史	否 =0，是 =1
糖尿病一级亲属家族史	否 =0，是 =1

表 3 Lasso 回归的系数和 λ min 值

Table 3 Coefficients and λ min values of Lasso regression

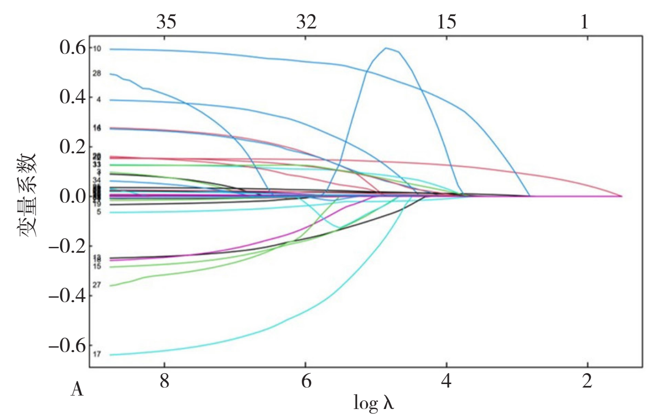
变量	系数	λ min
年龄	0.026 33	0.019 59
BMI	-0.001 48	
SBP	0.002 48	
吸烟	0.078 77	
饮酒	0.003 89	
高血压	0.000 02	
TC	0.000 50	
HDL-C	0.000 13	
CRP	0.001 84	
FPG	0.001 27	
ApoB	0.019 14	
LPA	0.000 16	
AST	0.000 79	
AST/ALT	0.004 44	
ACR	0.000 13	

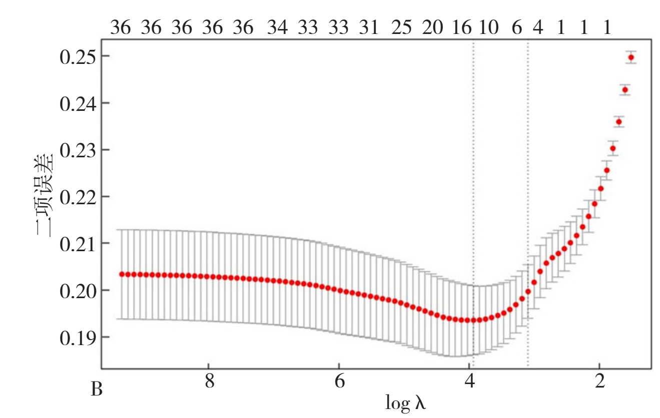
表 4 各模型在验证集中的判别性能

Table 4 Discriminative performance of each model on the validation set

模型	灵敏度	特异度	准确度	F1 值	AUC
Logistic 回归	0.693 9	0.800 0	0.750 0	0.723 4	0.836 7
RF	0.755 1	0.781 8	0.769 2	0.755 1	0.830 1
SVM	0.714 3	0.818 2	0.769 2	0.744 7	0.841 9
XGBoost	0.734 7	0.818 2	0.778 8	0.757 9	0.833 0
GBDT	0.755 1	0.836 4	0.798 1	0.778 9	0.834 9

注：RF= 随机森林，SVM= 支持向量机，XGBoost= 极端梯度增强，GBDT= 梯度增强决策树，AUC= 曲线下面积。





注：A 为 36 个变量的 Lasso 系数分布图；B 为 Lasso 回归模型中最佳参数（λ）的选择采用最低标准的 10 倍交叉验证，绘制二项偏差与 log λ 的关系曲线，在最小标准和最小标准的 1 标准差最优值处画虚线垂直线。

图 1 采用 Lasso 回归进行输入变量的筛选

Figure 1 Variable selection using lasso regression for input variables

注：RF= 随机森林，SVM= 支持向量机，XGBoost= 极端梯度增强，GBDT= 梯度增强决策树。

图 2 各模型在验证集中识别 CAS 的 ROC 曲线

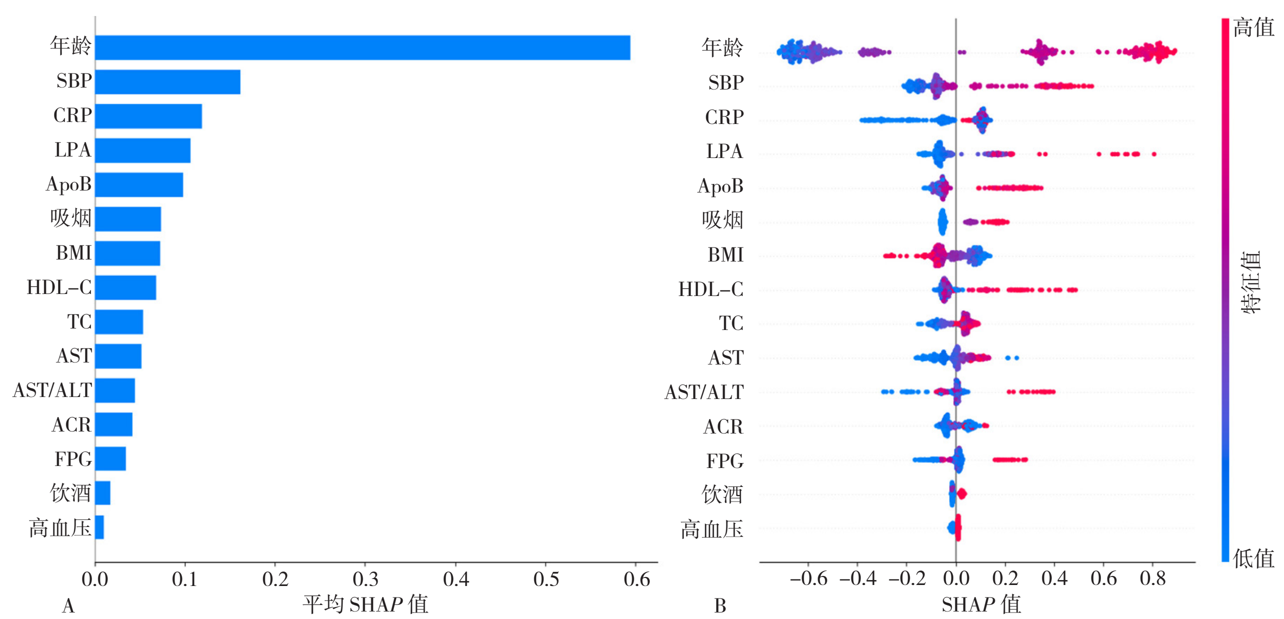
Figure 2 ROC curves of each model for identifying CAS on the validation set

响之间的关系。结果显示，变量重要性排序前 5 名依次为年龄、SBP、CRP、LPA、ApoB，图 3B 显示随着变量的升高而增加了 CAS 发生的风险。

3 讨论

本研究结果显示 FRS 中高风险群体中未识别出 CAS 的个体占比为 52.6%，与先前的研究结果相似[5，6]，提示根据 FRS 识别 CAS 准确性不足。为了提高 FRS 中高风险群体 CAS 早期识别的准确性，本研究构建了该群体 CAS 的风险预测模型，并筛选出最优模型，更准确地识别 CAS，以优化个体的预防和治疗策略，减轻医疗负担，避免医疗资源浪费。

本研究基于机器学习运用 Logistic 回归、RF、SVM、XGBoost 和 GBDT 算法构建了 5 个预测模型。全部模型的 AUC 值均较高，其中 GBDT 模型的综合判别效能最优（灵敏度 =0.755 1，特异度 =0.836 4，准确度 =0.798 1，F1 值 =0.778 9，AUC=0.834 9），与同类型的研究[21-23]相比，该模型被认为是具有较高精度的预测模型；在外部验证中也展现出了较强的泛化能力（AUC=0.794 0）。GBDT 算法是机器学习方法之一，也称为多元加性回归树，比 Logistic、决策树和 RF 算法具有更准确的识别能力和复杂的算法[24]，具有许多非线性变换和扎实的表现能力，不需要复杂的特征工程和变换[25]。GBDT 模型被广泛运用于疾病的识别，均表现出较好的判别性能。WU 等[21]运用 XGBoost、GBDT、RF 和 SVM 四种机器学习方法构建在无症状人群中颈动脉斑块识别模型，GBDT 模型 AUC 为 0.836 7，具有较高的判别性能。YE 等[26]利用重症监护医学信息数据库（MIMIC）Ⅳ数据库中患者的生命体征和实验室检查等多项指标，建立了基于机器学习的慢性肾脏疾病合并冠状动脉疾病的重症监护病房患者的住院死亡率的预测模型，其中最优模型为 GBDT 模型，AUC 可达 0.946。LIU 等[27]基于人工智能构建心肌梗死风险预测



注：A 为最优模型根据 SHAP 平均值排序的变量重要性图，B 为对变量重要性排序，并展示了变量对结局产生了何种影响；SBP= 收缩压，CRP=C- 反应蛋白，LPA= 脂蛋白 a，ApoB= 载脂蛋白 B，HDL-C= 高密度脂蛋白胆固醇，TC= 总胆固醇，AST= 天冬氨酸氨基转移酶，AST/ALT= 天冬氨酸氨基转移酶 / 丙氨酸氨基转移酶，ACR= 尿微量白蛋白肌酐比值，FPG= 空腹血糖。

图 3 最优模型可视化解释

Figure 3 Visualization of interpretation for the optimal model

模型，用于预警住院患者心肌梗死的发生，其中 GBDT 模型为最优模型，AUC 为 0.91。LIU 等[28]利用机器学习方法构建急性胰腺炎患者脓毒症风险预测模型，并将最优模型 GBDT 模型与 Logistic 回归模型和评分系统进行比较，显示判别性能优于 Logistic 回归模型和评分系统。SU 等[29]使用机器学习方法结合纵向数据来预测中国老年人 2 年内慢性肾脏疾病发展的风险，GBDT 模型表现出较好的判别性能。

本研究通过 SHAP 方法对 GBDT 模型进行可视化解释，对模型判别性能影响排名前 5 名的变量依次为年龄、SBP、CRP、LPA、ApoB，同时也表明年龄小、低 SBP、低 CRP、低 LPA 和低 ApoB 可以降低 CAS 发生的风险。张萍等人研究表明随着年龄增长，动脉管壁结构的胶原纤维和弹力纤维比例失调，导致动脉壁增厚、顺应性降低，加上一些疾病引起的血管内皮功能障碍和结构异常，促使粥样硬化的发生[30]。唐焱等[31]也发现年龄是颈动脉斑块形成的危险因素，随着年龄的增加 CAS 斑块也明显提升，并且不少研究也视其为独立危险因素。有研究表明高血压患者中 CAS 发生率更高，且 SBP 升高更为明显[32]。以往的研究表明，即使没有其他 CVD 危险因素存在，炎症仍然能够引发 CAS 的形成[33]。高水平的炎症可能导致内皮通透性的过度增加，这表示内皮屏障的完整性受到破坏。受损的内皮细胞通过进一步表达黏附分子和趋化因子，使白细胞能够在内皮上滚动、附着并最终进入血管壁，从而促进了血管壁炎症的发展[34]。研究表明，LPA 与颈动脉粥样硬化斑块发生关系密切，作用机理主要与胆固醇代谢以及纤维蛋白水解作用相关；高 LPA 患者心肌梗死和冠心病发病率高于健康人，脑动脉硬化患者 LPA 不仅显著高于健康人，还和病变的程度密切相关[35-36]。一项包括 8 项队列和 4 项病例对照研究的荟萃分析得出结论，ApoB 水平升高是首次缺血性卒中的危险因素[37]。本研究结果与上述研究结果一致，与临床实践也基本一致，说明本研究所构建的 GBDT 模型具有较强的合理性。

基层医疗卫生机构是实现当地群众就近就医、方便就医的首要环节，直接面对当地群众的医疗服务和卫生需求；同时，基层医疗卫生机构也是初级医疗卫生保健服务的主要提供者，发挥着医疗费用“守门人”和居民健康管理的重要作用，并向确有专科转诊需要的首诊患者提供专业性的建议[38]。有研究表明，心脑血管疾病患者的门诊治疗费用在家庭卫生支出中占比高达 44.05%，超过了所有疾病治疗费用在家庭卫生支出中所占的比例（34.85%），心脑血管疾病患者门诊治疗费用负担相对较为沉重，因此为了控制医疗费用和减轻疾病经济负担，有必要将心脑血管疾病列为未来疾病预防和控制的重点[39]。早诊早治是心脑血管疾病防治的关键，本研究所构建 CAS 风险预测模型的特征变量为公共卫生服务项目所包含的检测指标，容易获取，增加了模型的实用性，同时可以提高基层医疗人员识别 CAS 的简便性和准确性，这有助于早期识别并在病情恶化之前采取有效的预防和治疗策略，提高患者的生活质量，同时通过减少 CAS 引起的心血管事件，有望带来显著的社会经济效益，减轻医疗负担，提高健康资源的利用效率。

本研究存在一定的局限性：首先，采用方便抽样方法，存在着一定的选择偏倚；其次，女性占比偏高，可能与男性多在外地工作有关；再者，研究对象缺少相关服药情况，可能会对研究结果造成一定的影响；最后，研究对象大多来源于乡镇地区，对研究结果外推有一定的影响。

综上所述，本研究通过 Lasso 回归筛选出与 CAS 相关的特征变量，构建基于 Logistic 回归、RF、SVM、XGBoost 和 GBDT 的 FRS 中高风险群体 CAS 预测模型，通过灵敏度、特异度、准确度、F1 值和 AUC 值这 5 个评价指标综合评估判别性能，结果表明 GBDT 模型识别 CAS 的效果最佳，同时具有较强的泛化能力；运用 SHAP 方法对 GBDT 模型进行可视化解释，年龄、SBP、CRP、LPA、ApoB 是对模型判别效能最重要的变量，同时也是 CAS 的危险因素。这一研究成果有望帮助基层医疗人员进行更准确的评估，提高 CAS 的识别和治疗覆盖率，有助于合理分配医疗资源，并为 FRS 中高风险群体 CAS 的早期干预提供科学依据，进一步改善基层居民心血管健康、提高医疗服务水平以及促进社会公共卫生。在未来的研究和实践中，建议进一步验证和拓展模型的适用性，以确保其在不同人群中的有效性。

作者贡献：刘忠典、许琪、陈伊静、覃玲巧、陈淑萍、唐薇婷进行研究的实施、数据收集与整理；刘忠典负责进行统计学处理、结果的分析与解释及撰写论文；刘忠典、钟秋安进行论文的修订；钟秋安进行文章的构思与设计、可行性分析，负责文章的质量控制及审校。

本文无利益冲突。

刘忠典：https://orcid.org/0009-0003-3135-6800

参考文献

[1] 胡盛寿，王增武.《中国心血管健康与疾病报告 2022》概述[J]. 中国心血管病研究，2023，21（7）：577-600.

[2] SAKELLARIOS A I，BIZOPOULOS P，PAPAFAKLIS M I，et al. Natural history of carotid atherosclerosis in relation to the hemodynamic environment[J]. Angiology，2017，68（2）：109-118. DOI：10.1177/0003319716644138.

[3] JOHRI A M，NAMBI V，NAQVI T Z，et al. Recommendations for the assessment of carotid arterial plaque by ultrasound for the characterization of atherosclerosis and evaluation of cardiovascular risk：from the American society of echocardiography[J]. J Am

Soc Echocardiogr，2020，33（8）：917-933. DOI：10.1016/j.echo.2020.04.021.
［4］尤莉莉，陈新月，杨凌鹤，等.国家基本公共卫生服务项目十年评价（2009—2019年）系列报告（三）——国家基本公共卫生服务项目实施十年：挑战与建议［J］.中国全科医学，2022，25（26）：3221-3231. DOI：10.12114/j.issn.1007-9572.2022.0406.
［5］PEN A，YAM Y，CHEN L，et al. Discordance between Framingham Risk Score and atherosclerotic plaque burden［J］. Eur Heart J，2013，34（14）：1075-1082. DOI：10.1093/eurheartj/ehs473.
［6］易艳珊，农青娇，毛宝玉，等.基于弗明翰风险评分与血管内皮功能分类的心血管疾病危险因素研究［J］.中国全科医学，2018，21（16）：1959-1964. DOI：10.3969/j.issn.1007-9572.2018.16.011.
［7］RIDKER P M，BURING J E，RIFAI N，et al. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women：the Reynolds Risk Score［J］. JAMA，2007，297（6）：611-619. DOI：10.1001/jama.297.6.611.
［8］樊萌语，吕筠，何平平.国际体力活动问卷中体力活动水平的计算方法［J］.中华流行病学杂志，2014，35（8）：961-964. DOI：10.3760/cma.j.issn.0254-6450.2014.08.019.
［9］D'AGOSTINO R B Sr，VASAN R S，PENCINA M J，et al. General cardiovascular risk profile for use in primary care：the Framingham Heart Study［J］. Circulation，2008，117（6）：743-753. DOI：10.1161/CIRCULATIONAHA.107.699579.
［10］WANG X J，LI W Z，SONG F J，et al. Carotid atherosclerosis detected by ultrasonography：a national cross-sectional study［J］. J Am Heart Assoc，2018，7（8）：e008701. DOI：10.1161/JAHA.118.008701.
［11］陈润霖，何土凤，陶俐均，等.心血管危险因素对颈动脉内中膜进展的影响研究［J］.中国全科医学，2023，26（14）：1709-1715. DOI：10.12114/j.issn.1007-9572.2022.0750.
［12］TOUBOUL P J，HENNERICI M G，MEAIRS S，et al. Mannheim carotid intima-media thickness and plaque consensus（2004-2006-2011）. An update on behalf of the advisory board of the 3rd，4th and 5th watching the risk symposia，at the 13th，15th and 20th European Stroke Conferences，Mannheim，Germany，2004，Brussels，Belgium，2006，and Hamburg，Germany，2011［J］. Cerebrovasc Dis，2012，34（4）：290-296. DOI：10.1159/000343145.
［13］HORNE D J，CAMPO M，ORTIZ J R，et al. Association between smoking and latent tuberculosis in the U.S. population：an analysis of the National Health and Nutrition Examination Survey［J］. PLoS One，2012，7（11）：e49050. DOI：10.1371/journal.pone.0049050.
［14］KUO C C，WEAVER V，FADROWSKI J J，et al. Arsenic exposure，hyperuricemia，and gout in US adults［J］. Environ Int，2015，76：32-40. DOI：10.1016/j.envint.2014.11.015.
［15］LEVEY A S，STEVENS L A，SCHMID C H，et al. A new equation to estimate glomerular filtration rate［J］. Ann Intern Med，2009，150（9）：604-612. DOI：10.7326/0003-4819-150-9-200905050-00006.
［16］戴烨.基于《中国高血压防治指南（2018修订版）》对某院门诊降压药应用情况的调查［J］.中国社区医师，2022，38（12）：11-13.
［17］《中国老年型糖尿病防治临床指南》编写组.中国老年2型糖尿病防治临床指南（2022年版）［J］.中国糖尿病杂志，2022，30（1）：2-51. DOI：10.3969/j.issn.1006-6187.2022.01.002.
［18］EXPERT PANEL ON DETECTION E. Executive summary of the third report of the national cholesterol education program（NCEP）expert panel on detection，evaluation，and treatment of high blood cholesterol in adults（adult treatment panel Ⅲ）［J］. JAMA，2001，285（19）：2486-2497. DOI：10.1001/jama.285.19.2486.
［19］诸骏仁，高润霖，赵水平，等.中国成人血脂异常防治指南（2016年修订版）［J］.中国循环杂志，2016，31（10）：937-953.
［20］金文胜，潘长玉.国际糖尿病联盟关于代谢综合征定义的全球共识［J］.中华内分泌代谢杂志，2005，21（4）：附录4b-1-附录4b-2. DOI：10.3760/j.issn:1000-6699.2005.04.054.
［21］WU D，CUI G S，HUANG X X，et al. An accurate and explainable ensemble learning method for carotid plaque prediction in an asymptomatic population［J］. Comput Methods Programs Biomed，2022，221：106842. DOI：10.1016/j.cmpb.2022.106842.
［22］YU J，ZHOU Y，YANG Q，et al. Machine learning models for screening carotid atherosclerosis in asymptomatic adults［J］. Sci Rep，2021，11（1）：22236. DOI：10.1038/s41598-021-01456-3.
［23］龚军，钟小钢，谈军涛，等."网格搜索+XGBoost"算法建立儿童脓毒性休克预测模型［J］.解放军医学杂志，2020，45（12）：1270-1276.
［24］ZHOU Z H，FENG J. Deep forest［J］. Natl Sci Rev，2019，6（1）：74-86. DOI：10.1093/nsr/nwy108.
［25］ZHANG Z D，JUNG C. GBDT-MO：gradient-boosted decision trees for multiple outputs［J］. IEEE Trans Neural Netw Learn Syst，2021，32（7）：3156-3167. DOI：10.1109/TNNLS.2020.3009776.
［26］YE Z X，AN S Y，GAO Y X，et al. The prediction of in-hospital mortality in chronic kidney disease patients with coronary artery disease using machine learning models［J］. Eur J Med Res，2023，28（1）：33. DOI：10.1186/s40001-023-00995-x.
［27］LIU R，WANG M Y，ZHENG T，et al. An artificial intelligence-based risk prediction model of myocardial infarction［J］. BMC Bioinformatics，2022，23（1）：217. DOI：10.1186/s12859-022-04761-4.
［28］LIU F，YAO J，LIU C Y，et al. Construction and validation of machine learning models for sepsis prediction in patients with acute pancreatitis［J］. BMC Surg，2023，23（1）：267. DOI：10.1186/s12893-023-02151-y.
［29］SU D，ZHANG X Y，HE K，et al. Individualized prediction of chronic kidney disease for the elderly in longevity areas in China：machine learning approaches［J］. Front Public Health，2022，10：998549. DOI：10.3389/fpubh.2022.998549.
［30］张萍，郭秀丽，张鹏华.颈动脉粥样硬化与血管危险因素的相关性［J］.中国老年学杂志，2017，37（5）：1132-1134. DOI：10.3969/j.issn.1005-9202.2017.05.041.
［31］唐焱，周宏，罗光华，等.缺血性脑卒中患者CAS斑块超声、CT血管造影及临床相关危险因素分析［J］.中国动脉硬化杂志，

2016，24（4）：391-395.

[32] 高素颖，颜应琳，于凯，等. 急性缺血性脑卒中颈动脉粥样硬化的危险因素研究［J］. 中国全科医学，2021，24（3）：327-332. DOI：10.12114/j.issn.1007-9572.2020.00.401.

[33] TALEB S. Inflammation in atherosclerosis［J］. Arch Cardiovasc Dis，2016，109（12）：708-715. DOI：10.1016/j.acvd.2016.04.002.

[34] XU S W，ILYAS I，LITTLE P J，et al. Endothelial dysfunction in atherosclerotic cardiovascular diseases and beyond：from mechanism to pharmacotherapies［J］. Pharmacol Rev，2021，73（3）：924-967. DOI：10.1124/pharmrev.120.000096.

[35] 孔祥锋，王萍，陈明. 脂蛋白（a）与脑梗死患者颈动脉粥样硬化、纤维蛋白原、D-二聚体的关系［J］. 重庆医科大学学报，2011，36（9）：1101-1102. DOI：10.13406/j.cnki.cyxb.2011.09.027.

[36] 张玮，席艳，孙慧君，等. 脂蛋白A与血栓及动脉粥样硬化的关系［J］. 中国现代医学杂志，2007，17（20）：2500-2502，2505. DOI：10.3969/j.issn.1005-8982.2007.20.019.

[37] DONG H L，CHEN W，WANG X Y，et al. Apolipoprotein A1，B levels，and their ratio and the risk of a first stroke：a meta-analysis and case-control study［J］. Metab Brain Dis，2015，30（6）：1319-1330. DOI：10.1007/s11011-015-9732-7.

[38] 周忠良，范小静. 西部地区基层医疗卫生服务质量及提升策略［J］. 西安交通大学学报（社会科学版），2023，43（6）：188-200. DOI：10.15896/j.xjtuskxb.202306016.

[39] 张毓辉，翟铁民，柴培培，等. 我国心脑血管疾病治疗费用核算及预测研究［J］. 中国卫生经济，2019，38（5）：18-22. DOI：10.7664/CHE20190505.

（收稿日期：2024-02-19；修回日期：2024-04-30）
（本文编辑：康艳辉）